Robust Speaker Segmentation for Meetings: The ICSI Spring 2006 Diarization System

Xavier Anguera

Chuck Wooters

José Pardo

RT-06S Meeting Recognition Workshop

May 3rd, Bethesda, Maryland

Outline

- Tasks we participated in for RT-06s
- System description
- What's new since RT-05s
- Eval results
- Post-evaluation analysis
- Future work

Tasks and Submissions

(23 systems submitted)

- SPKR
  - Conf room:
    - MDM (1p, 4c)
    - SDM (1p, 1c)
  - Lecture room:
    - ADM (1p, 3c)
    - MDM (1p, 3c)
    - SDM (1p, 1c)
    - MSLA (1p)
- SAD
  - Conf room:
    - MDM (1p)
    - SDM (1p)
  - Lecture room:
    - ADM (1p)
    - MDM (1p)
    - SDM (1p)

p=primary, c=contrast

The ICSI Speaker Diarization System

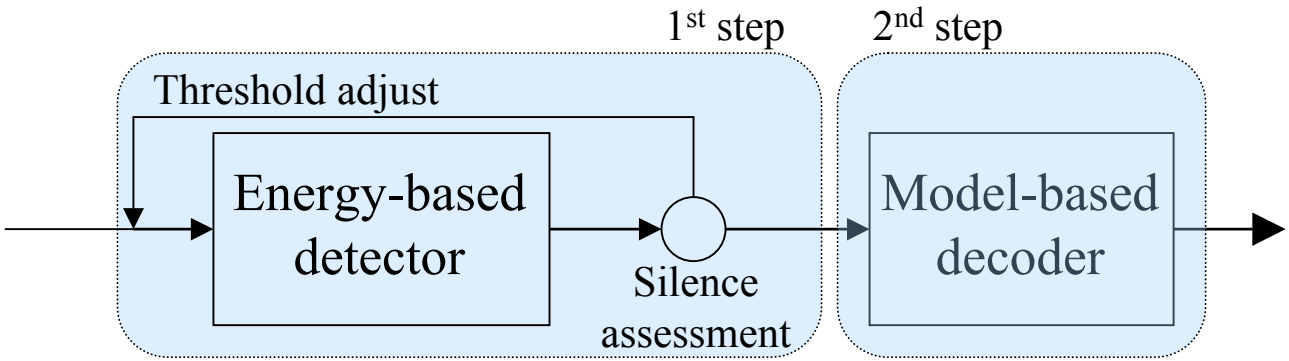
- Overall goal: Robustness and Portability across domains
- Agglomerative clustering
- Cluster merging uses a modified BIC
  - No penalty term in the BIC formula
- No pre-trained acoustic models
- Take advantage of multiple mics, when they are available, using delay&sum.

What's new?

- A new train-free speech/non-speech detector
- Improved cluster initialization ("friends and enemies" algorithm).
- Frame-based purification during merging.
- Initial number of clusters determined semi-automatically.
- Average speaker turn length only acoustically-driven.
- For MDM:
  - Modified delay-sum algorithm.
  - using delays in combination with acoustic features.

New Speech/non-speech

- Two-pass algorithm:
  - Step 1: Energy-based detector to locate silence regions with high confidence.
  - Step 2: Iterative clustering approach.
    - Models initialized with silence regions from first pass.
    - Convergence based on likelihood.

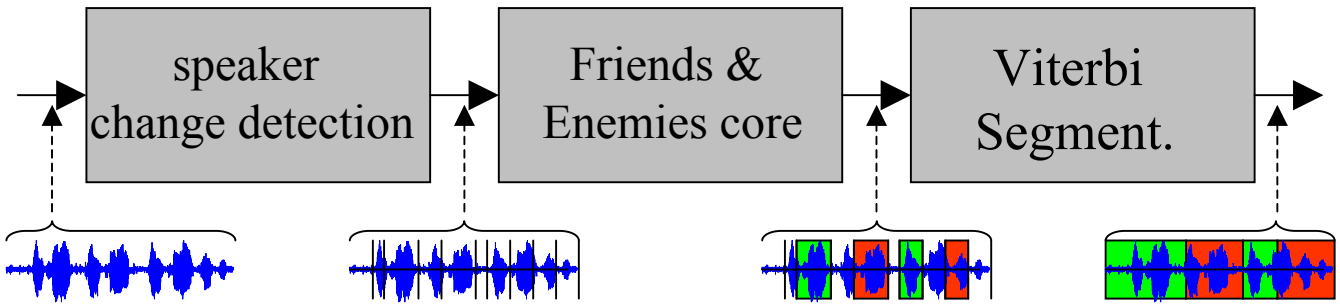


"**Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings**", X. Anguera et al., to appear Speaker Odyssey 2006

Improved Initialization

"Friends and enemies" algorithm:

1. Speaker change detection using BIC.
2. Create a first cluster as the most likely segment according to a general model.
3. Find the friends for this segment with highest cross-likelihood.
4. Iteratively find new "enemies" with smallest cross-lkld to existing clusters, and find its friends.
5. Train models and segment all the data.



"**Friends and Enemies: A Novel Initialization for Speaker Diarization**", X. Anguera et al., submitted to ICSLP 2006

Frame-based Purification

- Simpler than last year's segment-based purification.
- Problem: BIC-based cluster merging is adversely affected by "impure" data.
  - I.e. non-speech frames that occur in all clusters difficult to discriminate between clusters.
- Observation: most non-speech frames obtain the best lkld in all models.
- Proposed solution: Try to detect "impure" frames, using a likelihood-based metric, and exclude these during cluster comparison.

$$\overline{\mathcal{L}}(x[i] \mid \Theta_A) = \frac{1}{Q} \sum_{j=-(Q/2)}^{(Q/2)-1} \sum_{m=1}^{\widetilde{M}} \log(W_A[m] \mathcal{N}_{A,m}(x[i+j]))$$

"**Frame Purification for Cluster Comparison in Speaker Diarization**", X. Anguera et al., MMUA 2006. And a related paper at ICASSP06

Initial Number of Clusters

- Last year an informed guess was made about the number of initial clusters (10 for Conf., 5 for Lect.).
- This year the initial number of clusters (K_{init}) is determined based on the length of the meeting (N_{total}), using:

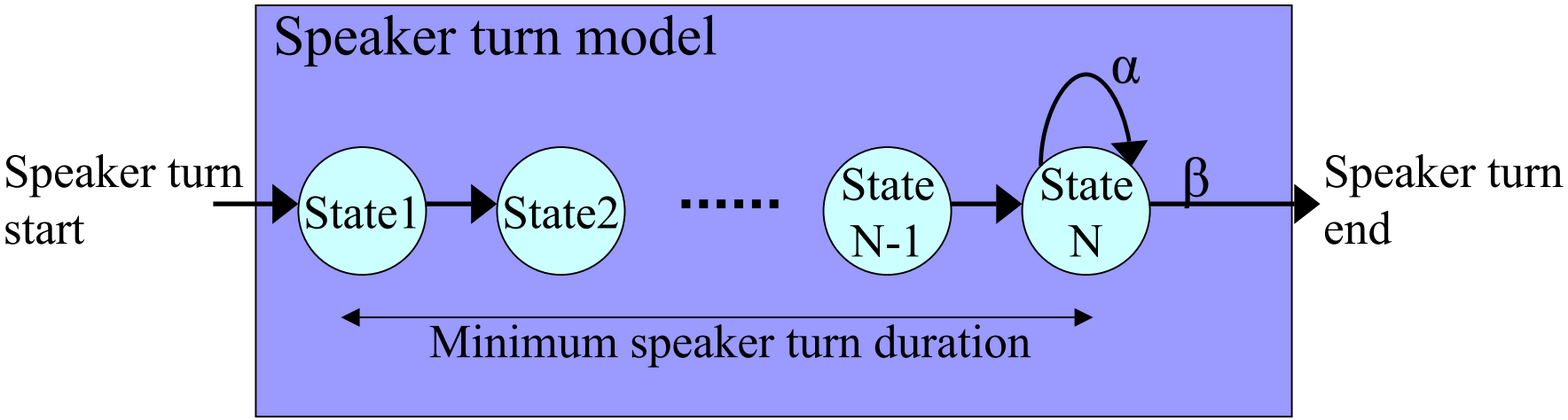
$$K_{init} = \frac{N_{total}}{GM_{clus} CCR}$$

GM_{clus}= # Gauss per cluster, CCR = # frames necessary to train a Gaussian, to be optimized.

"Automatic Cluster complexity and Quantity Selection: Towards Robust Speaker Diarization", X. Anguera et al., MLMI 2006

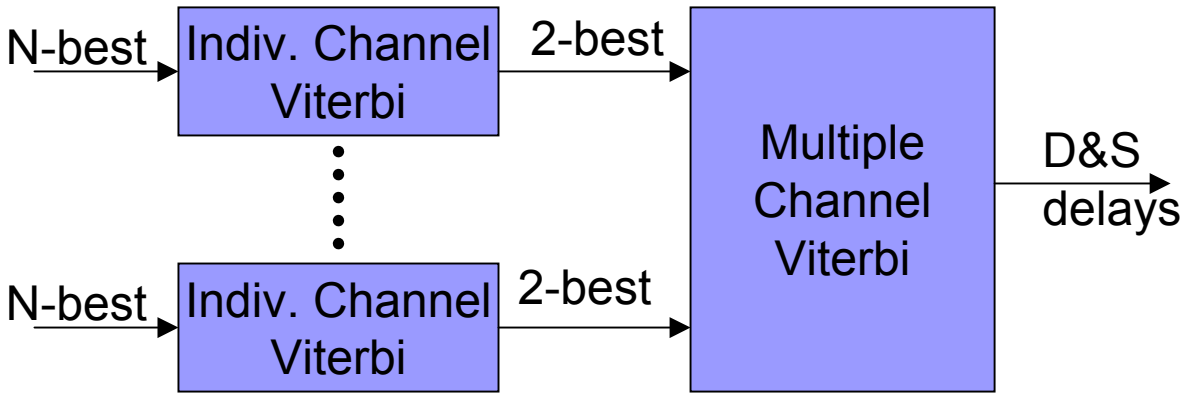
Modified average speaker turn length

- The average speaker turn duration was artificially constrained by the chosen minimum turn duration and parameters α and β.
- Making α = 1 and β = 1 makes the average turn duration dependent only on the acoustic data.



MDM-specific Improvements

- Modified delay-sum algorithm (new code)
  - New noise thresholding to eliminate “bad quality” delays.
  - Smoothing of delays using Viterbi in 2 steps:
    1. Select the 2-best delays from N-best GCC-PHAT peaks in each individual channel.
    2. Select best delays vector across all channels.



MDM-specific Improvements (II)

- Modified delay-sum algorithm (II)
  - Modified channel weighting (α = 0.95)

$$W_i[n] = \alpha W_i[n-1] + (1-\alpha) Xc_i[n]$$

Before:

$$Xc_i[n] = \frac{Xcorr(i, ref)[n]}{\sum_j Xcorr(j, ref)[n]}$$

Now:

$$Xc_i[n] = \frac{\sum_k Xcorr(i,k)[n]}{\sum_j \sum_k Xcorr(j,k)}$$

  - "Suspicious" frames elimination if $Xc_i[n] < \frac{1}{3(Nc-1)}$

MDM-specific Improvements (III)

- Using delay&sum delays as features in combination with acoustic features.
  - Models built using a weighted combination of two separate feature streams, modeled by 2 separate GMM models.

$$\log p(x_{ac}[i], x_{del}[i] \mid \Theta_{tot}) = \alpha \cdot \log p(x_{ac}[i] \mid \Theta_{ac}) + (1-\alpha) \cdot \log p(x_{del}[i] \mid \Theta_{dels})$$

  - It is used in the Viterbi segmentation and in the BIC models comparison.

$$\Delta BIC_{tot} = \alpha \cdot \Delta BIC_{aco} + (1-\alpha) \cdot \Delta BIC_{del}$$

"Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features and Inter-Channel Time Differences", J. M. Pardo et al.,submitted to ICSLP 2006

Eval Results
Conference Room - Spkr

Cond.	*System ID*	*%DER*	*Description*
MDM	**p-wdels**	**35.77**	Primary system.
	c-newspnspdelay	35.77	Same as last year, w/ delays and new spnsp.
	c-wdelsfix	38.26	Same as primary, but init clusts = 16.
	c-nodels	41.93	Same as primary, but no delay feats.
	c-oldbase	42.36	Same as last year's system, w/ new spnsp.
SDM	**p-nodels**	**43.59**	Primary system (no delay feats).
	c-oldbase	43.93	Same as last year's system /w new spnsp.

Breakdown of errors by type for **primary systems**:

	Miss	*FA*	*SpNsp*	*Spkr*	*Total*
MDM	27.60	1.10	28.70	7.20	35.77
SDM	28.90	0.80	29.70	13.90	43.59

Eval Results
Lecture Room - Spkr

Cond.	*System ID*	*%DER*	*Description*
MDM	**p-wdels**	**24.01**	Primary system. (Same as conf. room system)
	c-nodels	23.63	Same as primary, but no delay feats.
	c-wdelsfix	24.53	Same as primary, but init # clusts = 10.
	c-guessone	26.96	Guess one speaker all the time. No sp/nsp.
SDM	**p-nodels**	**23.95**	Primary system (no delay feats).
	c-guessone	26.96	Guess one speaker all the time. No sp/nsp.
ADM	**p-wdels**	**21.05**	Same as MDM primary, but using all channels.
	c-nodels	20.24	Same as primary, but no delay feats.
	c-wdelsfix	22.11	Same as primary, but init # clusts = 10.
	c-guessone	26.96	Guess one speaker all the time. No sp/nsp.
MSLA	**p-guessone**	**26.96**	Guess one speaker all the time. No sp/nsp.

Eval Results
Conference Room - SAD

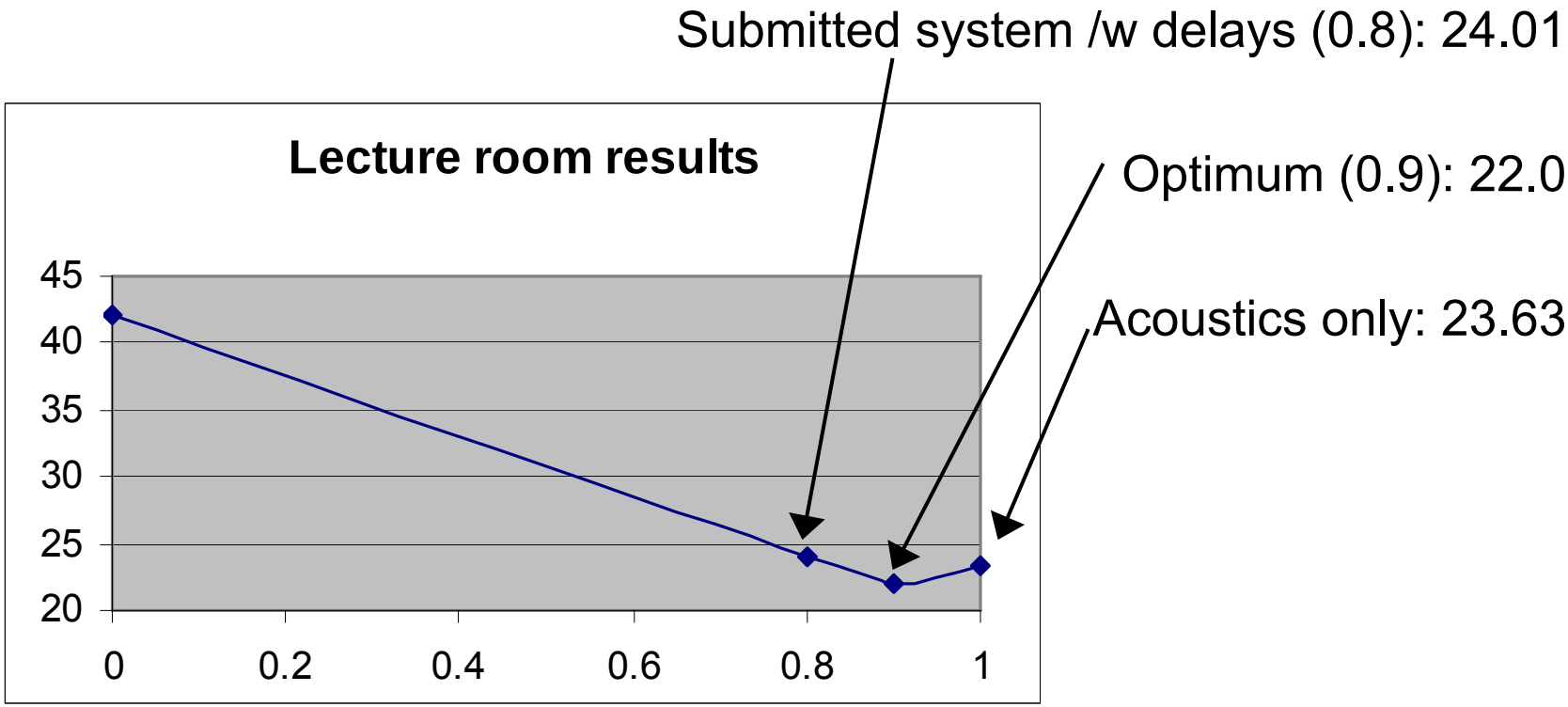
Cond.	*System ID*	*%Error*	*%Miss*	*%FA*	*Description*
MDM	**p-dual**	**23.51**	22.76	0.8	Two-pass system tuned to forced alignments.
SDM	**p-dual**	**24.95**	24.24	0.8	Two-pass system tuned to forced alignments.

Eval Results
Lecture Room - SAD

Cond.	***System ID***	***%Error***	***%Miss***	***%FA***	***Description***
MDM	**p-dual**	**13.83**	9.3	4.5	Two-pass system tuned to forced alignments.
SDM	**p-dual**	**14.59**	10.0	4.6	Two-pass system tuned to forced alignments.
ADM	**p-dual**	**13.22**	9.3	3.9	Two-pass system tuned to forced alignments.

Post-evaluation analysis

- Lecture room



Post-evaluation analysis

- Conference room data evaluated using ICSI-SRI Forced Alignments.

Cond.	***System ID***	***%DER FA***	***%DER hand***
MDM	**p-wdels**	**19.16**	**35.77**
	c-newspnspdelay	20.03	35.77
	c-wdelsfix	23.32	38.26
	c-nodels	27.46	41.93
	c-oldbase	27.01	42.36
SDM	**p-nodels**	**28.25**	**43.59**
	c-oldbase	28.21	43.93

Future work

- Continue work on cluster initialization and purification.
- Add other acoustic features (e.g. PLP, prosody, etc.) as additional streams.
  - Try to improve both SDM and MDM/ADM conditions.
  - Dynamic stream weights.
- Experiment with alternatives to speech/non-speech, e.g. voiced/unvoiced.

Questions?